

**UNIVERSIDAD EAFIT**  
**ST1800/ST1801 ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN**  
**2025-1**

**Nombre:**

Jose Luis Bedoya Martinez, [jlbedoyam@eafit.edu.co](mailto:jlbedoyam@eafit.edu.co)  
Álvaro Javier Mutis Guerrero, [ajmutisg@eafit.edu.co](mailto:ajmutisg@eafit.edu.co)  
Juan Luis Amaya Arbeláez, [ilamayaa@eafit.edu.co](mailto:ilamayaa@eafit.edu.co)  
Kevin Genez Valencia, [kgenezv@eafit.edu.co](mailto:kgenezv@eafit.edu.co)

Enlace para acceder a toda la documentación en GitHub:

<https://github.com/kdgenz/ARI-T1-P2.git>

## Contenido

Parte 1: .....	3
1. LABORATORIO 0:.....	3
2. LABORATORIO 1:.....	7
3. LABORATORIO 2:.....	10
4. LABORATORIO 3:.....	14
LAB 3.1: REDSHIFT .....	14
LAB 3.2: REDSHIFT SPECTRUM.....	16
5. LABORATORIO 4:.....	19
Parte 2: .....	22

Ilustración 1 Activar la cuenta en AWS ACADEMY .....	3
Ilustración 2 Iniciar Laboratorio .....	4
Ilustración 3. Descargar el archivo de credenciales .PEM para conexión remota por SSH .....	5
Ilustración 4. Crear una máquina virtual sencilla Ubuntu 22.04 .....	6
Ilustración 5. verificar que todo funciona correctamente .....	7
Ilustración 6. Crear Nuevo Bucket.....	8
Ilustración 7. Crear Folder .....	9
Ilustración 8. Copiar Archivos a Bucket.....	9
Ilustración 9. Ingreso a AWS Glue .....	10
Ilustración 10. Set Crawlers .....	10
Ilustración 11. Conectar a S3 .....	10
Ilustración 12. Run Crawler .....	11
Ilustración 13: Catalogación con AWS Glue.....	11
Ilustración 14. configuración Directorio Athena .....	12
Ilustración 15. Implementación DWH Tickit .....	13

Ilustración 16. Algunas consultas .....	14
Ilustración 17. Creación Cluster .....	14
Ilustración 18. configuración Cluster.....	15
Ilustración 19. Consultas en Redshift .....	16
Ilustración 20. Crear un rol IAM para Amazon Redshift .....	17
Ilustración 21. Crear tabla externa .....	17
Ilustración 22. Crear una tabla con datos externos en S3 .....	18
Ilustración 23.Crear una tabla nativa en redshit.....	18
Ilustración 24. Cargar datos en la table ‘event2.....	19
Ilustración 25. Descargar datos y Almacenar en S3 .....	19
Ilustración 26. Copiar datos desde S3 .....	20
Ilustración 27. Crear Modelo .....	20
Ilustración 28. Verificar estado del modelo .....	20
Ilustración 29. Predicciones del modelo .....	21

**Parte 1:****1. LABORATORIO 0:**

*Activar la cuenta y el curso en AWS ACADEMY*

Welcome Aboard!

In order to finish signing you up for the course **AWS Academy Learner Lab [110512]**, we'll need a little more information.

Login:

Password: \*

Time Zone:

Yes, I'd like Canvas to provide my contact information to [Amazon Web Services](#) (AWS) so AWS can share the latest news about AWS services and related offerings with me by email, post or telephone.

Personal Email:   
Get the latest news and offers from AWS Academy.

You may unsubscribe from receiving AWS news and offers at any time by following the instructions in the communications received. AWS handles your information as described in the [AWS Privacy Notice](#). Providing Canvas with your information may involve transferring it to another country. For questions about how Canvas will handle your information, please contact Canvas directly or refer to its privacy policy.

I agree to the [Canvas Instructure Acceptable Use Policy](#) and to the [AWS Learner Terms and Conditions](#). The information you provide will be handled by AWS as described in the [AWS Privacy Notice](#).

**Register**

*Ilustración 1 Activar la cuenta en AWS ACADEMY*

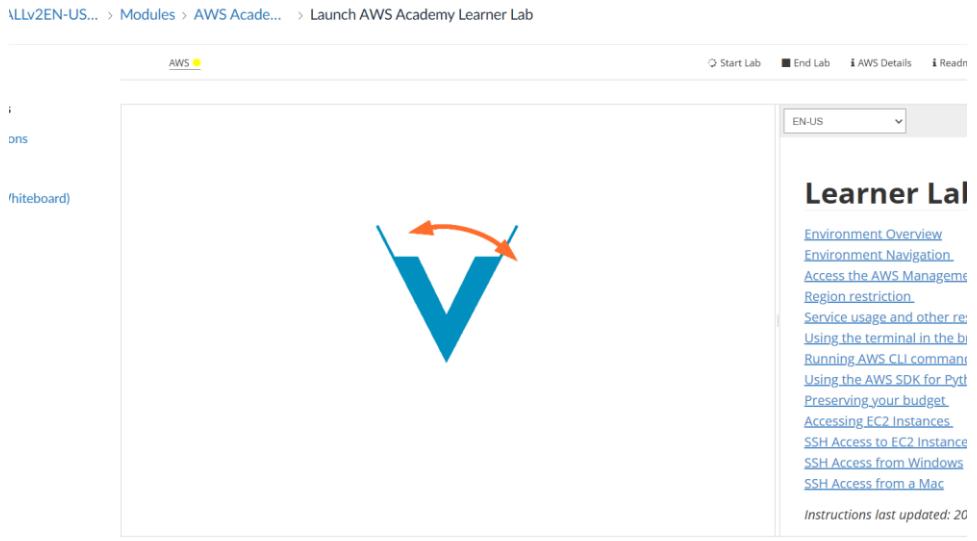
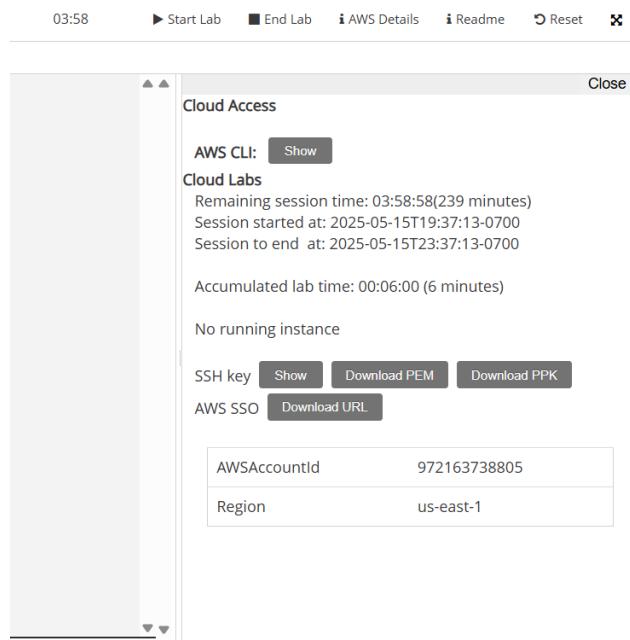


Ilustración 2 Iniciar Laboratorio



*Ilustración 3. Descargar el archivo de credenciales. PEM para conexión remota por SSH*

The first screenshot shows a file list with 'Nombre' and 'Fecha' columns. It includes a file named 'labsuser.pem' under the 'hoy' folder, with a download icon and the date '15/0'.

The second screenshot shows a file download dialog for 'vockey.pem' with the download date '15/05/2025 9:40 p'.

The third screenshot shows the AWS Applications console. It has a header with icons for search, notifications, and help, and dropdowns for 'Estados Unidos (Norte de Virginia)' and 'voclabs/user4087263=jbedoyam@eafit.edu.co @ 9721-6373-8805'. Below the header are buttons for 'Restablecer al diseño predeterminado' and '+ Agregar widgets'. The main area shows 'Aplicaciones (0) Información' with a 'Crear aplicación' button. It includes a 'Seleccionar región' dropdown set to 'us-east-1 (Región actual)', a search bar 'Buscar aplicaciones', and a navigation bar with page number '1'. At the bottom, there are filters for 'Nombre', 'Descripción', 'Región', and 'Cuenta o'.

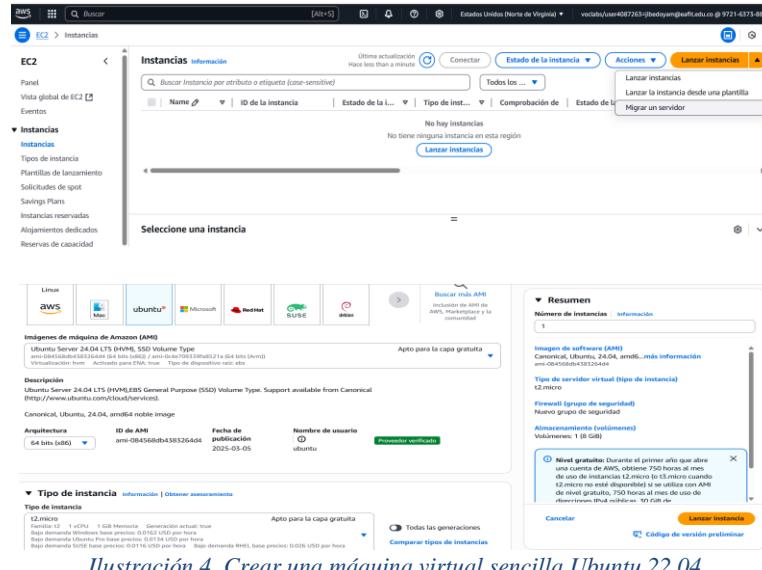
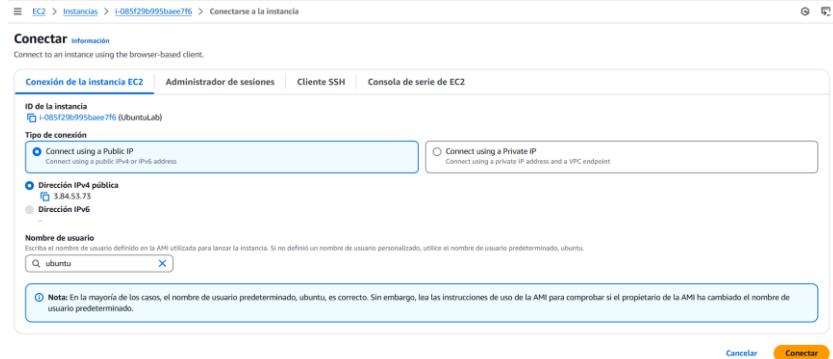
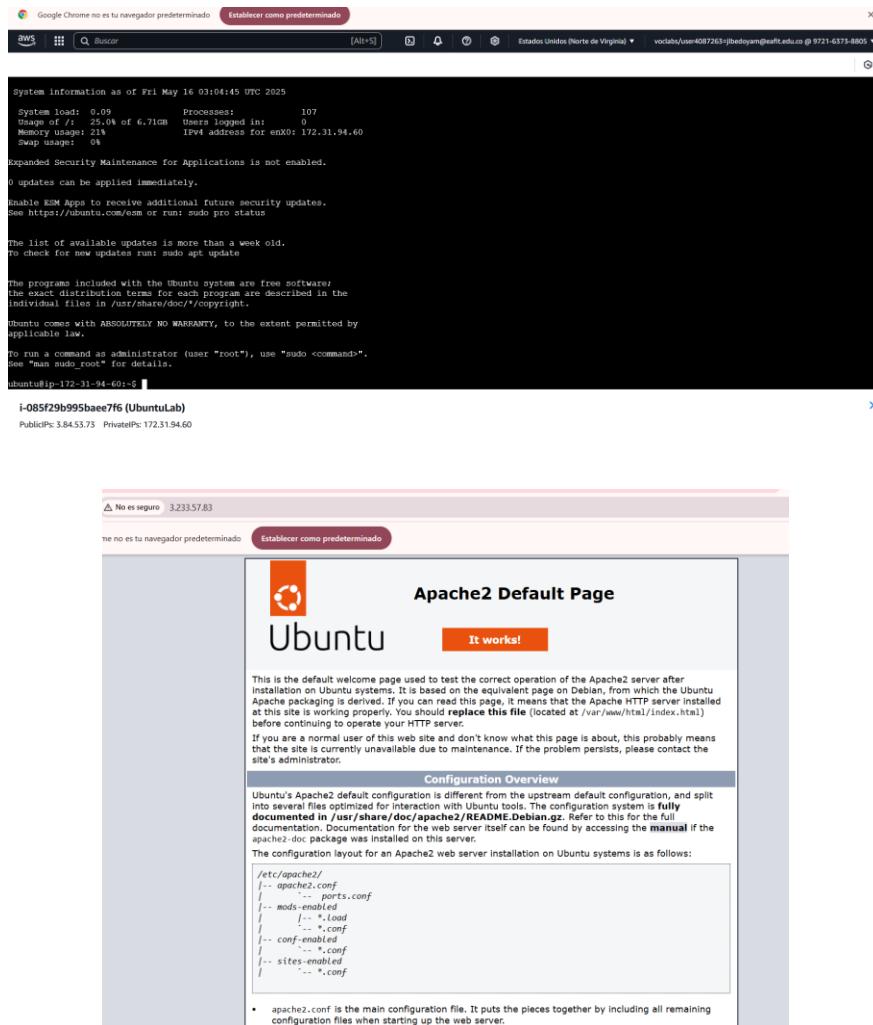


Ilustración 4. Crear una máquina virtual sencilla Ubuntu 22.04





*Ilustración 5. verificar que todo funciona correctamente*

## 2. LABORATORIO 1:

ALMACENAMIENTO DE DATOS EN AWS S3

The screenshot shows the 'Create a bucket' wizard on the Amazon S3 service. The first step, 'Bucket type', is displayed. It offers two options: 'General purpose' (selected) and 'Directory'. Both options are described with their respective features. Below this, there's a section for 'Bucket name' with a placeholder 'jbedoyam101'. A note states that bucket names must be 3 to 63 characters and unique. There are also sections for 'Object Ownership' (set to 'Bucket owner enforced') and 'Block Public Access settings for this bucket' (with several checkboxes for different access control options). At the bottom, a note about turning off public access is shown.

The second part of the screenshot shows the 'Success' step of the 'Create New Bucket' wizard. It displays a message: 'Successfully created bucket "jbedoyam101". To upload files and folders, or to configure additional bucket settings, choose View details.' Below this, there's an 'Account snapshot - updated every 24 hours' section and a 'General purpose buckets' table. The table lists one bucket: 'jbedoyam101' (General purpose bucket, US East (N. Virginia) us-east-1 region, IAM Access Analyzer enabled, Creation date May 17, 2025, 19:16:31 (UTC-05:00)).

Ilustración 6. Crear Nuevo Bucket

**Create folder**

Use folders to group objects in buckets. When you create a folder, S3 creates an object using the name that you specify followed by a slash (/). This object then appears as folder on the console. [Learn more](#)

Your bucket policy might block folder creation  
If your bucket policy prevents uploading objects without specific tags, metadata, or access control list (ACL) grants, you will not be able to create a folder using this configuration. Instead, you can use the [upload configuration](#) to upload an empty folder and specify the appropriate settings.

**Folder name**  
datasets2

**Server-side encryption** Info  
Server-side encryption protects data at rest.

The following encryption settings apply only to the folder object and not to sub-folder objects.

**Server-side encryption**

**Don't use encryption key**  
The bucket setting for default encryption are used to encrypt the folder object when storing it in Amazon S3.

**Specify an encryption key**  
This option allows you to use your own encryption key to encrypt the folder object before storing it in Amazon S3.

If your bucket policy requires objects to be encrypted with a specific encryption key, you must specify the same encryption key when you create a folder. Otherwise, folder creation will fail.

**Create folder**

Ilustración 7. Crear Folder

Upload status																																																																																																																																																																																																																												
<a href="#">View details</a>																																																																																																																																																																																																																												
After you navigate away from this page, the following information is no longer available.																																																																																																																																																																																																																												
<b>Summary</b>																																																																																																																																																																																																																												
Destination	s3://jlbedoymlab1/datasets2/	Succeeded	40 files, 154.7 MB (100.00%)	Failed	0 files, 0 B (0%)																																																																																																																																																																																																																							
<a href="#">View details</a>																																																																																																																																																																																																																												
<b>Files and Folders</b> (40 total, 154.7 MB)																																																																																																																																																																																																																												
<table border="1"> <thead> <tr> <th>Name</th> <th>Type</th> <th>Size</th> <th>Status</th> <th>Error</th> </tr> </thead> <tbody> <tr><td>parsons.csv</td><td>text/csv</td><td>761.8 kB</td><td></td><td></td></tr> <tr><td>shakespeare</td><td>dataset/</td><td>74.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny.csv</td><td>dataset/</td><td>28.2 MB</td><td></td><td></td></tr> <tr><td>tiny_tiny2.csv.zip</td><td>dataset/</td><td>23.0 MB</td><td></td><td></td></tr> <tr><td>tiny_tiny3.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny4.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny5.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny6.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny7.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny8.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny9.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny10.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny11.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny12.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny13.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny14.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny15.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny16.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny17.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny18.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny19.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny20.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny21.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny22.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny23.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny24.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny25.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny26.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny27.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny28.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny29.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny30.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny31.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny32.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny33.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny34.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny35.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny36.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny37.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny38.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny39.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> <tr><td>tiny_tiny40.csv</td><td>dataset/</td><td>1.0 kB</td><td></td><td></td></tr> </tbody> </table>						Name	Type	Size	Status	Error	parsons.csv	text/csv	761.8 kB			shakespeare	dataset/	74.0 kB			tiny_tiny.csv	dataset/	28.2 MB			tiny_tiny2.csv.zip	dataset/	23.0 MB			tiny_tiny3.csv	dataset/	1.0 kB			tiny_tiny4.csv	dataset/	1.0 kB			tiny_tiny5.csv	dataset/	1.0 kB			tiny_tiny6.csv	dataset/	1.0 kB			tiny_tiny7.csv	dataset/	1.0 kB			tiny_tiny8.csv	dataset/	1.0 kB			tiny_tiny9.csv	dataset/	1.0 kB			tiny_tiny10.csv	dataset/	1.0 kB			tiny_tiny11.csv	dataset/	1.0 kB			tiny_tiny12.csv	dataset/	1.0 kB			tiny_tiny13.csv	dataset/	1.0 kB			tiny_tiny14.csv	dataset/	1.0 kB			tiny_tiny15.csv	dataset/	1.0 kB			tiny_tiny16.csv	dataset/	1.0 kB			tiny_tiny17.csv	dataset/	1.0 kB			tiny_tiny18.csv	dataset/	1.0 kB			tiny_tiny19.csv	dataset/	1.0 kB			tiny_tiny20.csv	dataset/	1.0 kB			tiny_tiny21.csv	dataset/	1.0 kB			tiny_tiny22.csv	dataset/	1.0 kB			tiny_tiny23.csv	dataset/	1.0 kB			tiny_tiny24.csv	dataset/	1.0 kB			tiny_tiny25.csv	dataset/	1.0 kB			tiny_tiny26.csv	dataset/	1.0 kB			tiny_tiny27.csv	dataset/	1.0 kB			tiny_tiny28.csv	dataset/	1.0 kB			tiny_tiny29.csv	dataset/	1.0 kB			tiny_tiny30.csv	dataset/	1.0 kB			tiny_tiny31.csv	dataset/	1.0 kB			tiny_tiny32.csv	dataset/	1.0 kB			tiny_tiny33.csv	dataset/	1.0 kB			tiny_tiny34.csv	dataset/	1.0 kB			tiny_tiny35.csv	dataset/	1.0 kB			tiny_tiny36.csv	dataset/	1.0 kB			tiny_tiny37.csv	dataset/	1.0 kB			tiny_tiny38.csv	dataset/	1.0 kB			tiny_tiny39.csv	dataset/	1.0 kB			tiny_tiny40.csv	dataset/	1.0 kB		
Name	Type	Size	Status	Error																																																																																																																																																																																																																								
parsons.csv	text/csv	761.8 kB																																																																																																																																																																																																																										
shakespeare	dataset/	74.0 kB																																																																																																																																																																																																																										
tiny_tiny.csv	dataset/	28.2 MB																																																																																																																																																																																																																										
tiny_tiny2.csv.zip	dataset/	23.0 MB																																																																																																																																																																																																																										
tiny_tiny3.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny4.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny5.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny6.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny7.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny8.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny9.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny10.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny11.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny12.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny13.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny14.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny15.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny16.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny17.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny18.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny19.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny20.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny21.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny22.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny23.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny24.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny25.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny26.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny27.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny28.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny29.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny30.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny31.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny32.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny33.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny34.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny35.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny36.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny37.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny38.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny39.csv	dataset/	1.0 kB																																																																																																																																																																																																																										
tiny_tiny40.csv	dataset/	1.0 kB																																																																																																																																																																																																																										

**Objects** (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

**Actions**

Copy S3 link Copy URL Download Open Delete Actions Create folder Upload

**Objects** (1/2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

**Actions**

Copy S3 link Copy URL Download Open Delete Actions Create folder Upload

Name	Type	Last modified	Size	Storage class
datasets2	Folder	-	-	-
tiny_tiny2	Folder	-	-	-

**Objects** (1/2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

**Actions**

Copy S3 link Copy URL Download Open Delete Actions Create folder Upload

Name	Type	Last modified	Size	Storage class
datasets2	Folder	-	-	-
tiny_tiny2	Folder	-	-	-
tiny_tiny3	Folder	-	-	-

Ilustración 8. Copiar Archivos a Bucket

s3://jlbedoymlab1/datasets/

### 3. LABORATORIO 2:

#### IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS ATHENA.

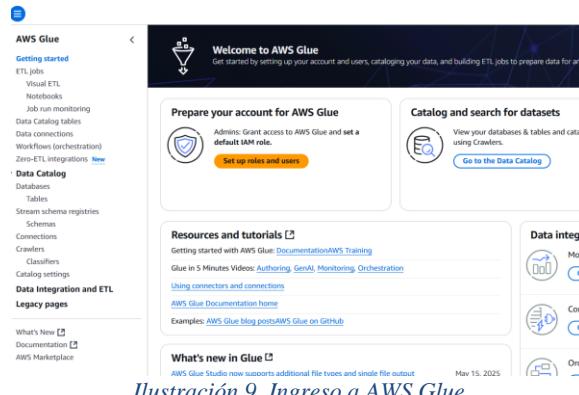


Ilustración 9. Ingreso a AWS Glue

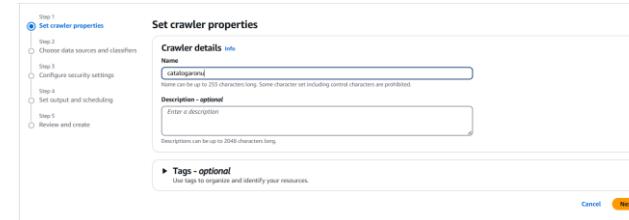


Ilustración 10. Set Crawlers



**Add data source**

**Data source**  
Choose the source of data to be crawled.

**Network connection - optional**  
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

**Location of S3 data**  
 In this account  
 In a different account

**S3 path**  
Browse for or enter an existing S3 path.  
    
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

**Subsequent crawler runs**  
This field is a global field that affects all S3 data sources.  
 Crawl all sub-folders  
Crawl all folders again with every subsequent crawl.  
 Crawl new sub-folders only  
Only crawl S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.  
 Crawl based on events  
Rely on Amazon S3 events to control what folders to crawl.  
 Sample only a subset of files  
 Exclude files matching pattern

Ilustración 11. Conectar a S3

The screenshot shows the 'Configure security settings' step of a crawler creation process. It includes sections for IAM role (choosing 'Lambda'), Lake Formation configuration (optional), and security configuration (optional). Below this, a green success message states: 'One crawler successfully created. The following crawler is now created: "catalogaroru"'. The crawler properties section shows the name 'catalogaroru', IAM role 'Lambda', database 'latidb', state 'READY', and table prefix 'Table prefix'. The 'Crawler runs' tab is empty.

Ilustración 12. Run Crawler

The screenshot shows the 'Database properties' for 'labsdb'. It lists the database name, description, location, and creation date (May 16, 2023 at 14:27:12). The 'Tables (2)' tab shows two tables: 'export' and 'hdi'. Both tables are located in 's3://jbedoymlab1/onsu' and are in CSV format. The 'Add tables using crawler' button is visible.

Edit schema: hdi

The screenshot shows the 'Edit schema' interface for the 'hdi' table. It displays 9 columns with the following details:

#	Column name	Data type	Partition key	Comment
1	id	bigint	-	-
2	country	string	-	-
3	human development index (hdi)	double	-	-
4	leb	double	-	-
5	mean years of schooling	double	-	-
6	expected years of schooling	double	-	-
7	gross national income (gni) per ca...	bigint	-	-
8	gni per capita rank minus hdi rank	bigint	-	-
9	nonincome hdi	double	-	-

Buttons for 'Delete', 'Edit', and 'Add' are at the top right, and 'Save as new table version' is at the bottom right.

Ilustración 13: Catalogación con AWS Glue

The image shows two screenshots of the Amazon Athena interface. The top screenshot is the 'Query editor' showing the 'Data' sidebar with 'AwsDataCatalog' selected as the data source, 'None' as the catalog, and 'labsdb' as the database. The 'Tables and views' section lists 'Tables (2)' with 'export' and 'hd1' selected, and 'Views (0)'. The main area shows 'Query 1' with the SQL command 'SELECT \* FROM "labsdb"."hd1" limit 10;'. The bottom screenshot is the 'Manage settings' page under 'Query result location and encryption'. It shows the 'Location of query result - optional' field set to 's3://jbedoymlab1/athena'. Other settings include 'Lifecycle configuration' (disabled), 'Expected bucket owner - optional' (disabled), 'Assign bucket owner full control over query results' (unchecked), and 'Encrypt query results' (unchecked). A note at the top of the settings page states: 'Athena now supports typeahead code suggestions to speed up SQL query development. Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.'

*Ilustración 14. configuración Directorio Athena*

This screenshot shows the 'Query editor' with two queries open. 'Query 1' contains the SQL command 'SELECT \* FROM "labsdb"."hd1" limit 10;'. 'Query 2' is currently active and contains the SQL command '1 | SELECT \* FROM "labsdb"."hd1" limit 10;'. The 'Data' sidebar is identical to the previous screenshot, showing 'AwsDataCatalog' as the data source, 'None' as the catalog, and 'labsdb' as the database. The 'Tables and views' section shows 'Tables (2)' with 'export' and 'hd1' selected. The bottom of the screen shows 'Query results' and 'Query stats' buttons.

*Crear base de datos ticket (ejercicio propuesto)*

The screenshot shows the 'Add data source' dialog box. In the 'Data source' section, 'S3' is selected from a dropdown menu. Below it, the 'Network connection - optional' section is shown with a note about network connections. The 'Location of S3 data' section has 'In this account' selected. The 'S3 path' section shows a path 's3://jlbedoymlab1/ticketdb2'. The 'Subsequent crawler runs' section contains several options: 'Crawl all sub-folders' (selected), 'Crawl new sub-folders only', 'Crawl based on events', 'Sample only a subset of files', and 'Exclude files matching pattern'. At the bottom right are 'Cancel' and 'Add an S3 data source' buttons.

This screenshot shows the 'Set output and scheduling' step. It includes sections for 'Output configuration' (targeting 'dticket'), 'Table name prefix' (set to 'ticket'), 'Maximum table threshold' (set to '0'), and 'Advanced options'. The 'Crawler schedule' section shows 'Frequency' set to 'On demand'. Navigation buttons 'Previous' and 'Next' are at the bottom right.

The top part shows the 'Crawler properties' for 'dticket', including fields like Name, IAM role, Security configuration, Database, State, and Table prefix. The bottom part shows the 'Tables' section with a table of 7 tables: category, data, events, listings, sales, users, and venue, each with their respective database and location details.

Ilustración 15. Implementación DWH Tickit

SQL Ln 1, Col 1

Run again Explain Cancel Create

Ilustración 16. Algunas consultas

## 4. LABORATORIO 3:

*IMPLEMENTACIÓN DE UN DATA WAREHOUSE CON AWS REDSHIFT y REDSHIFT SPECTRUM*

### LAB 3.1: REDSHIFT

Cluster configuration

Cluster identifier  
This is the unique key that identifies a cluster:

Choose the size of the cluster  
 I'll choose  
 Help me choose

Node type [Info](#)  
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

Number of nodes  
Enter the number of nodes that you need.  
  
Range (1-32)

Configuration summary [Info](#)  
dc2.large | 1 node

<b>\$182.50/month</b> Estimated on-demand compute price Save more than 60% of your costs by purchasing reserved nodes. <a href="#">Learn more about pricing</a>	<b>160 GB</b> Total compressed storage The total storage capacity for the cluster if you deploy the number of nodes that you chose.
---	---

Ilustración 17. Creación Cluster

Database configurations

Admin user name  
Enter a login ID for the admin user of your DB instance.

Admin password  
Select how to manage your admin password.  
 Manage admin credentials in AWS Secrets Manager [Info](#)  
AWS manages a KMS key that encrypts your data.  
 Generate a password [Info](#)  
Amazon generates a random password for your admin password.  
 Manually add the admin password [Info](#)  
Manually enter the admin password.  
  
Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except '.', ',', or ';'.

Show password

St1800232.

**Associated IAM roles (2/2) info**

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.

IAM roles	Status	Role type
LabRole	Not applied	--
myRedshiftRole	Not applied	--

**additional configurations** Use defaults

These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

The default configuration has been updated to **default.redshift-2.0**. This configuration enforces SSL connections for enhanced security.

**Network**  
Using default VPC (vpc-044ad25b00fb73656) and default subnet.

**Security**  
Using default (sg-0febe3c3c1ea5708) cluster security group.

**Configuration**  
Using defaultredshift-2.0 parameter group with no database encryption.

**Backup**  
Automated snapshots are created about every eight hours or following every 5 GB per node of data changes, whichever comes first.

**Maintenance**  
Using current maintenance track.

**Create cluster**

**Clusters (1) info**

Cluster	Status	Cluster namespace	Average query dur...	Average number of ...	Availability Zone	Multi-AZ	Storage capacity us...
redshift-cluster-1 (d2.large)   1 node   160 GB	Available	a72c9005-cb41-4186-...	~	~	us-east-1b	No	0 %

Ilustración 18. configuración Cluster

**Redshift query editor v2**

**Sales per event**

```

1 SET search_path to tickit;
2 SELECT eventid, total_price
3   FROM (SELECT eventid, total_price, percentile(1000) over(order by total_price desc) as percentile
4          FROM (SELECT eventid, sum(pricetpaid) total_price
5                 FROM tickit.sales
6                GROUP BY eventid) Q, tickit.event
7            WHERE Q.eventid = E.eventid
8            AND percentile = 1
9        ORDER BY total_price desc);

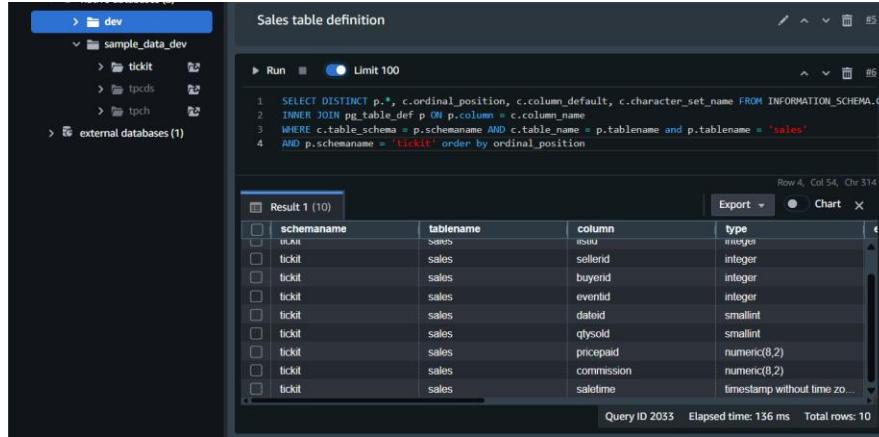
```

**Total quantity per buyer**

```

1 SELECT firstname, lastname, total_quantity
2   FROM (SELECT buyerid, sum(qtysold) total_quantity
3          FROM tickit.sales
4            GROUP BY buyerid
5            ORDER BY total_quantity desc limit 10) Q, tickit.users
6      WHERE Q.buyerid = user.id
7      ORDER BY Q.total_quantity desc;

```



The screenshot shows a Redshift query editor interface. On the left, a sidebar displays database structures: 'sample\_data\_dev' contains 'ticket', 'tpcds', and 'tpch'; 'external databases (1)' contains 'ticket'. The main area is titled 'Sales table definition' and contains a SQL query:

```

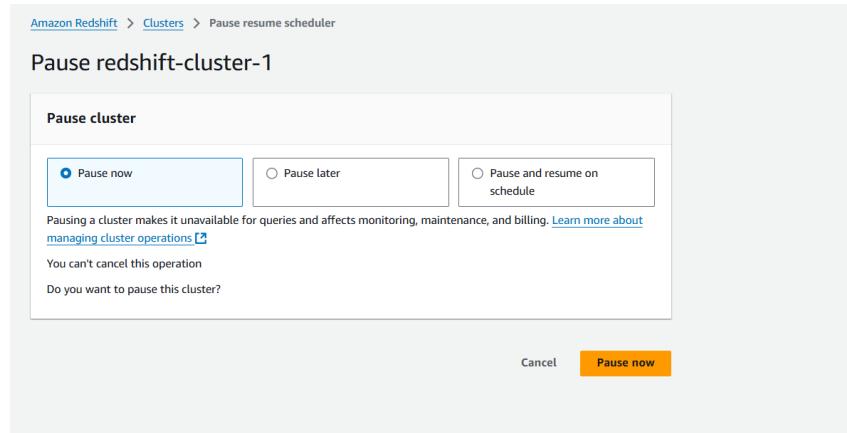
1 SELECT DISTINCT p.* , c.ordinal_position, c.column_default, c.character_set_name FROM INFORMATION_SCHEMA.COLUMNS p
2 INNER JOIN pg_table_def p ON p.column = c.column_name
3 WHERE c.table_schema = p.schemaname AND c.table_name = p.tablename AND p.tablename = 'sales'
4 AND p.schemaname = 'ticket' ORDER BY ordinal_position
    
```

The results table shows 10 rows of data:

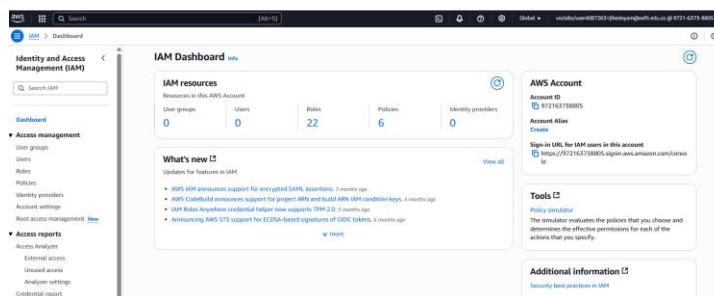
schemaname	tablename	column	type
ticket	Sales	issbu	integer
ticket	Sales	sellerid	integer
ticket	Sales	buyend	integer
ticket	Sales	eventid	integer
ticket	Sales	dateid	smallint
ticket	Sales	qtsold	smallint
ticket	Sales	pricepaid	numeric(8,2)
ticket	Sales	commission	numeric(8,2)
ticket	Sales	saletime	timestamp without time zone

Query ID: 2033 Elapsed time: 136 ms Total rows: 10

Ilustración 19. Consultas en Redshift



## LAB 3.2: RDSHIFT SPECTRUM



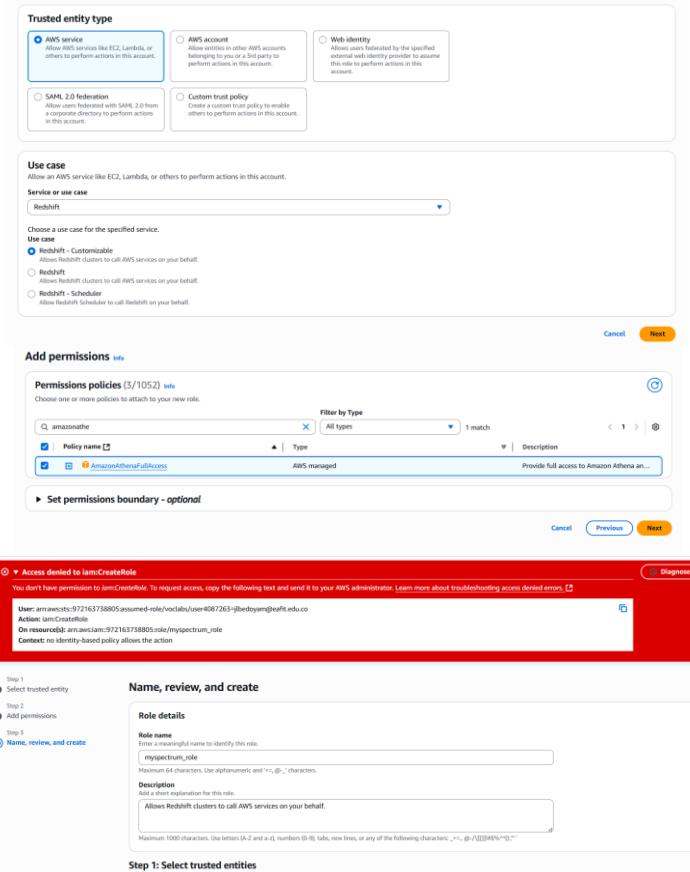


Ilustración 20. Crear un rol IAM para Amazon Redshift

arn:aws:iam::972163738805:role/LabRole

```

Run Explain Limit 100 Isolated session redshift-clust... dev Schedule
1 create external schema myspectrum_schema
2 from data catalog;
3 database myspectrum_db;
4 iam_role 'arn:aws:iam::972163738805:role/LabRole';
5 create external database if not exists;

```

Result 1

Summary

Info: External database "myspectrum\_db" created

Ilustración 21. Crear tabla externa

```

1  create external table myspectrum_schema.sales(
2    salesid integer,
3    listid integer,
4    sellerid integer,
5    buyerid integer,
6    eventid integer,
7    dateid smallint,
8    qtysold smallint,
9    pricepaid decimal(8,2),
10   commission decimal(8,2),
11   saletime timestamp)
12  row format delimited
13  fields terminated by '\t'
14  stored as textfile
15  location 's3://jlbodoymlab1/datasets/sales/'
16  table properties ('numRows='172000');

```

Row 15, Col 44, Chr 400

**Result 1**

Result set query:

```

/* RQEY2-RwCE2JFdxP */
create external table myspectrum_schema.sales(
salesid integer,
listid integer,
sellerid integer,
buyerid integer,
eventid integer,
dateid smallint,
qtysold smallint,
pricepaid decimal(8,2),
commission decimal(8,2),
saletime timestamp)
row format delimited
fields terminated by '\t'
stored as textfile
location 's3://jlbodoymlab1/datasets/sales/'

```

Ilustración 22. Crear una tabla con datos externos en S3

```

1  create table event2(
2    eventid integer not null distkey,
3    venueid smallint not null,
4    catid smallint not null,
5    dateid smallint not null sortkey,
6    eventname varchar(200),
7    starttime timestamp);

```

Row 7, Col 22, Chr 200

**Result 1**

**Summary**

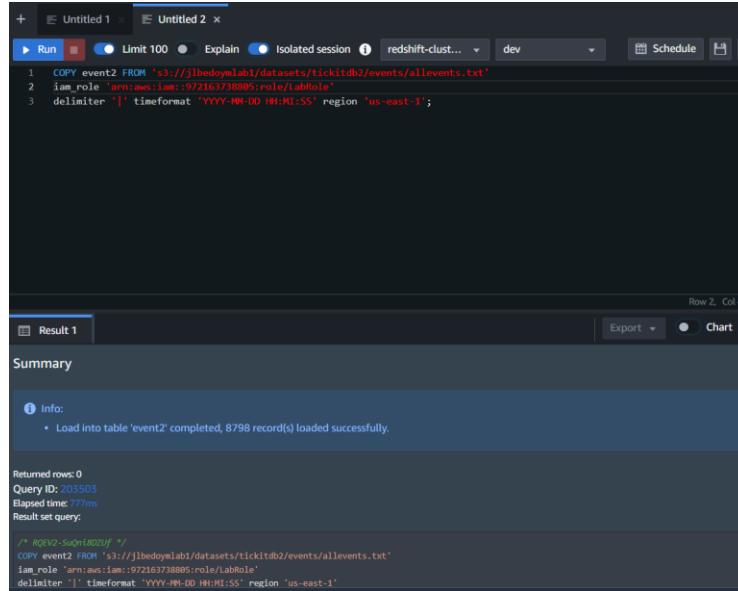
Returned rows: 0  
Query ID: 200419  
Elapsed time: 3ms  
Result set query:

```

/* RQEY2-RwCE2JFdxP */
create table event2(
eventid integer not null distkey,
venueid smallint not null,
catid smallint not null,
dateid smallint not null sortkey,
eventname varchar(200),
starttime timestamp)

```

Ilustración 23. Crear una tabla nativa en redshit



The screenshot shows the AWS Redshift console interface. A query is being run in a session titled 'Untitled 2'.

```

1 COPY event2 FROM 's3://jlbedoymlab1/datasets/tickitdb2/events/allevnts.txt'
2 iam_role 'arn:aws:iam::972163738805:role/labRole'
3 delimiter '|' timeformat 'YYYY-MM-DD HH:MI:SS' region 'us-east-1';
    
```

The 'Result 1' tab is selected, displaying the following summary:

- Info:**
  - Load into table 'event2' completed, 8798 record(s) loaded successfully.
- Returned rows: 0**
- Query ID:** 203503
- Elapsed time:** 777ms
- Result set query:**

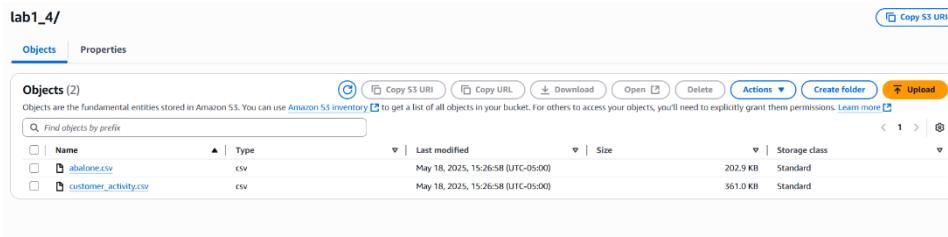
```

/* R02V2-SgQn1B02Uf */
COPY event2 FROM 's3://jlbedoymlab1/datasets/tickitdb2/events/allevnts.txt'
iam_role 'arn:aws:iam::972163738805:role/labRole'
delimiter '|' timeformat 'YYYY-MM-DD HH:MI:SS' region 'us-east-1';
    
```

Ilustración 24. Cargar datos en la table 'event2'

## 5. LABORATORIO 4:

AWS REDSHIFT con Aprendizaje de Máquina.

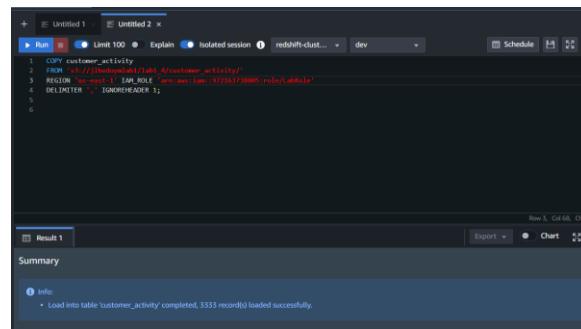


The screenshot shows the AWS S3 console for a bucket named 'lab1\_4/'. The 'Objects' tab is selected, displaying the following contents:

Name	Type	Last modified	Size	Storage class
abalone.csv	csv	May 18, 2025, 15:26:58 (UTC-05:00)	202.9 KB	Standard
customer_activity.csv	csv	May 18, 2025, 15:26:58 (UTC-05:00)	561.0 KB	Standard

Ilustración 25. Descargar datos y Almacenar en S3

s3://jlbedoymlab1/lab1\_4/



The screenshot shows the AWS Redshift console interface. A query is being run in a session titled 'Untitled 2'.

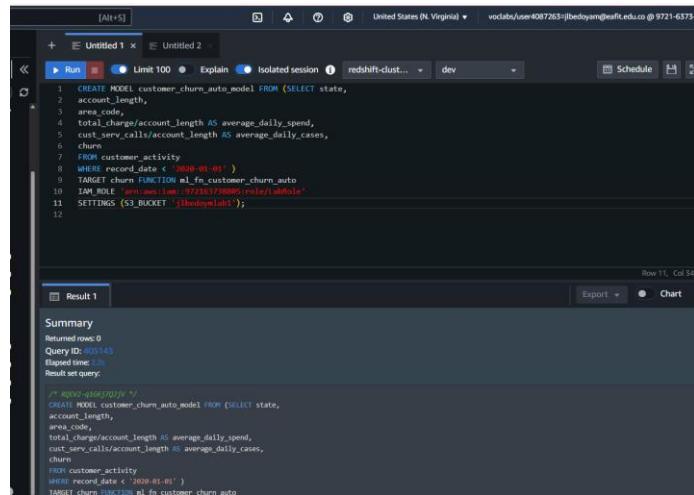
```

1 COPY customer_activity
2 FROM 's3://jlbedoymlab1/lab1_4/customer_activity/'
3 iam_role 'arn:aws:iam::972163738805:role/labRole'
4 DELIMITER ',' IGNOREHEADER 1;
5
    
```

The 'Result 1' tab is selected, displaying the following summary:

- Info:**
  - Load into table 'customer\_activity' completed, 3553 record(s) loaded successfully.

Ilustración 26. Copiar datos desde S3



The screenshot shows the AWS Redshift console interface. A query named 'Untitled 1' is displayed in the editor pane:

```

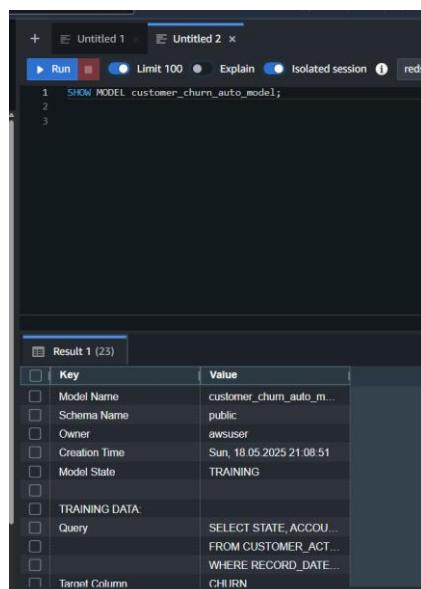
1 CREATE MODEL customer_churn_auto_model FROM (SELECT state,
2 account_length,
3 area_code,
4 total_charge/account_length AS average_daily_spend,
5 cust_serv_calls/account_length AS average_daily_cases,
6 churn
7 FROM customer_activity
8 WHERE record_date < '2000-01-01')
9 TARGET churn FUNCTION ml_fn_customer_churn_auto
10 IAM_ROLE 'arn:aws:iam::07236730085:role/redshift'
11 SETTINGS ($S3_BUCKET 'jibedoyimlah!');
12

```

The 'Result 1' pane shows the summary of the executed command:

Summary  
Returned rows: 0  
Query ID: 2023-05-18T10:00:45  
Elapsed time: 2.03s  
Result set query:  
`/* RDS-0161 */  
CREATE MODEL customer_churn_auto_model FROM (SELECT state,  
account_length,  
area_code,  
total_charge/account_length AS average_daily_spend,  
cust_serv_calls/account_length AS average_daily_cases,  
churn  
FROM customer_activity  
WHERE record_date < '2000-01-01')  
TARGET churn FUNCTION ml_fn_customer_churn_auto`

Ilustración 27. Crear Modelo



The screenshot shows the AWS Redshift console interface. A query named 'Untitled 1' is displayed in the editor pane:

```

1 SHOW MODEL customer_churn_auto_model;
2
3

```

The 'Result 1' pane displays the details of the created model:

Key	Value
Model Name	customer_chum_auto_m...
Schema Name	public
Owner	awsuser
Creation Time	Sun, 18 05 2025 21:08:51
Model State	TRAINING
TRAINING DATA:	
Query	<code>SELECT STATE, ACCOU...</code>
	<code>FROM CUSTOMER_ACT...</code>
	<code>WHERE RECORD_DATE...</code>
Target Column	CHURN

Ilustración 28. Verificar estado del modelo

The screenshot shows a Redshift Jupyter Notebook interface. At the top, there are two tabs: 'Untitled 1' and 'Untitled 2'. Below the tabs, there are several buttons: 'Run', 'Limit 100', 'Explain', 'Isolated session', and dropdown menus for 'redshift-clust...', 'dev', and 'United States (N. Virginia)'. The code in 'Untitled 1' is:

```

1 SELECT phone,
2     e1_fn_customer_churn_auto(
3         state,
4         account_length,
5         area_code,
6         total_charge/account_length ,
7         cust_serv_calls/account.length )
8     AS active FROM customer_activity WHERE record_date > '2020-01-01';
9

```

The 'Result 1 (100)' section displays a table with two columns: 'phone' and 'active'. The data is as follows:

	phone	active
0	382-4657	True.
0	358-1921	False.
0	330-8626	False.
0	329-9001	False.
0	335-4719	True.
0	330-8173	True.
0	351-7269	True.
0	350-8884	False.
0	393-7984	False.
0	343-4696	False.
0	418-6412	False.

The screenshot shows a Redshift Jupyter Notebook interface. The code in the cell is a complex SQL query involving multiple joins and conditions, including subqueries and window functions. The 'Result 2 (21)' section displays a table with five columns: 'state', 'total\_per\_state', 'nonchurners', 'churners', and 'ratio'. The data is as follows:

state	total_per_state	nonchurners	churners	ratio
AK	38	28	5	5.6
AL	44	44	4	11.0
AR	35	35	3	11.7
AS	35	35	3	11.7
DC	30	30	3	10.0
DE	45	45	8	5.6
DC	33	33	3	11.0
FL	34	34	8	4.25
GA	33	33	5	6.6

Ilustración 29. Predicciones del modelo

**Parte 2:****Trabajo 1 – Diseño e implementación de una solución analítica básica en una arquitectura Batch en AWS.**

Desarrollar el proyecto basado en un ciclo de vida clásico de la Arquitectura Batch:

- 1) Fuentes de datos: Datos del PROYECTO INTEGRADOR.
- 2) Ingesta de los datos: copia de archivos desde la interfaz web de AWS S3 o desde EMR/S3.
- 3) Almacenamiento: almacenar los datos originales en la zona raw, los datos preparados en la zona ‘trusted’ y los datos analizados, así como modelos en la zona ‘refined’.
- 4) Procesamiento de datos:
  - a. La preparación de datos, la puede hacer utilizando dataframes de Spark, y deberá colocar los datos preparados en la zona ‘trusted’.
  - b. Catalogación de datos para EDA con SparkSQL: Realizar la catalogación de datos (creación de los esquemas) con una de 2 herramientas: AWS Glue (automática) o con EMR/Hive (manual). La base de datos para las tablas generadas deberá llamarse ‘proyecto1db’
  - c. Realizar Análisis Exploratorio de Datos con SparkSQL, utilizando como entrada las tablas de la base de datos ‘proyecto1db’, los datos de salida deben ser almacenados en la zona ‘refined’ del datalake/s3.
  - d. Diseñar e implementar un modelo básico de Aprendizaje Automático de la librería SparkML o SparkGraphX. Los datos de salida deben ser almacenados en la zona ‘refined’ del datalake/s3.
  - e. Aplicación: inicialmente realizar algún esquema de visualización de datos, conectándose con la zona ‘refined’, considerada la zona de salida / acceso, para exponer los resultados del proyecto. Cada equipo podría creativamente definir algún otro mecanismo de aplicación en complemento de la visualización (si aplica).

## Desarrollo del trabajo

### Creación S3 y conexión al cluster

The screenshot shows the AWS S3 console interface. On the left, there's a sidebar with options like 'Amazon S3', 'Storage Lens', and 'CloudShell'. The main area displays a summary message: 'Instantánea de la cuenta: actualizada cada 24 horas' (Todas las regiones de AWS). Below this, there are two tabs: 'Buckets de uso general' (selected) and 'Buckets de directorio'. A search bar and filter options ('Nombre', 'Región de AWS', 'Analizador de acceso de IAM', 'Fecha de creación') are present. Two buckets are listed:

Nombre	Región de AWS	Analizador de acceso de IAM	Fecha de creación
aws-logs-851725502060-us-east-1	EE.UU. Este (Norte de Virginia) us-east-1	<a href="#">Ver analizador para us-east-1</a>	22 May 2025 5:50:33 PM -05
proyectointegrador1017	EE.UU. Este (Norte de Virginia) us-east-1	<a href="#">Ver analizador para us-east-1</a>	21 May 2025 5:52:36 PM -05

### Configuración del cluster

The screenshot shows the 'Crear clúster' (Create Cluster) wizard in the AWS EMR console. The left panel contains several sections:

- Paquete de aplicaciones:** Options include Spark Interactive, Core Hadoop, Flink, HBase, Presto, Trino, and Custom. Under Custom, Flink 1.20.0 is selected.
- Configuración del Catálogo de datos de AWS Glue:** Options include using the AWS Glue catalog for external metadata and applying automatic updates.
- Opciones del sistema operativo:** Options include Version de Amazon Linux (selected), Imagen de máquina de Amazon (AMI) personalizada, and Aplicar automáticamente las actualizaciones más recientes de Amazon Linux.

The right panel displays the 'Resumen' (Summary) and 'Información' (Information) section, which includes:

- Nombre y aplicaciones - obligatorio:** Version de Amazon EMR: emr-7.8.0.
- Paquete de aplicaciones:** Custom (Flink 1.20.0, HCatalog 3.1.3, Hadoop 3.4.1, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, Phoenix 5.2.1, Presto 0.287, Spark 3.5.4, Tez 0.10.2, Trino 467).
- Configuración del clúster - obligatorio:** Grupos de instancias uniformes: Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge).
- Aprovisionamiento y escalado de clústeres - obligatorio:** Configuration of provisioning: Número de núcleos: 1 instancia, Tamaño de la tarea: 1 instancia.
- Redes - obligatorio:** Configuration of networking.

## Configuración del cluster

## Creación del cluster

## Zona raw, zona trusted y zona refined

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
athena-results/	Carpeta	-	-	-
dashboard/	Carpeta	-	-	-
Imagenes/	Carpeta	-	-	-
ipyter/	Carpeta	-	-	-
raw/	Carpeta	-	-	-
refined/	Carpeta	-	-	-
script/	Carpeta	-	-	-
trusted_\$folder\$	-	26 May 2025 11:58:19 AM -05	0 B	Estándar
trusted/	Carpeta	-	-	-

Se define la ruta “proyectointegrador1017” en S3 la zona raw para cargar tres archivos CSV en DataFrames:

- BasePrediccion\_28092021.csv: Contiene los datos a utilizar para el entrenamiento del modelo.
- DiccionarioInicial.csv: Contiene un par de variables para la traducción de palabras similares y la corrección de errores de tipo al ingresar la observación.
- StopWords.csv: Contiene un listado de palabras a ser eliminadas del texto.

Con esto se constituye la lectura de datos crudos que serán transformados. Posteriormente, se define manualmente un esquema con 30 columnas para asegurar los tipos de variables y que sean consistentes para una vez más leer el archivo BasePrediccion\_28092021, pero esta vez con los tipos de variables deseados para evitar errores. Además, es necesario cargar desde S3 la lista de StopWords que serán utilizadas para eliminar el ruido del texto de las observaciones.

The screenshot shows the AWS S3 console interface. The left sidebar lists buckets under 'Amazon S3' and 'Storage Lens'. The main area displays the contents of the 'raw/' bucket. There are four objects listed:

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
Base_Prediccion_28092021.csv	CSV	23 May 2025 9:48:49 PM -05	2.5 MB	Estandar
DiccionarioIncial.csv	CSV	22 May 2025 9:48:49 PM -05	54.3 KB	Estandar
Stopwords.csv	CSV	25 May 2025 4:12:32 PM -05	13.9 KB	Estandar
Variables_seleccionadas.csv	CSV	22 May 2025 9:48:49 PM -05	1.1 KB	Estandar

Una vez se tiene todo cargado se comienza con la limpieza y depuración del campo observación. Se comienza eliminando el encabezado que contiene información redundante y luego se limpia este fragmento de texto, eliminando caracteres no alfanuméricos, secuencia de números, sellos identificadores e información administrativa de cada registro. Se registran las UDF para cada caso de limpieza y se aplican para generar el df\_limpio.

Después, con df\_limpio, se toma la lista de stopwords y se crea la UDF para dejar solo las palabras relevantes. También, para normalizar la semántica se aplica un diccionario de sinónimos para estandarizar el vocabulario de las observaciones. A esto se le añadió la codificación de la variable objetivo “Causa Evento” en tres clases: normal, anomalía o no revisado. El resultado de esto es una tabla estructurada con variables categóricas, numéricas y textuales limpias, almacenada en la zona trusted.

The screenshot shows the AWS S3 console interface. The left sidebar lists buckets under 'Amazon S3' and 'Storage Lens'. The main area displays the contents of the 'trusted/' bucket. There is one object listed:

Nombre	Tipo
Base_Entrenamiento.CSV	Carpeta

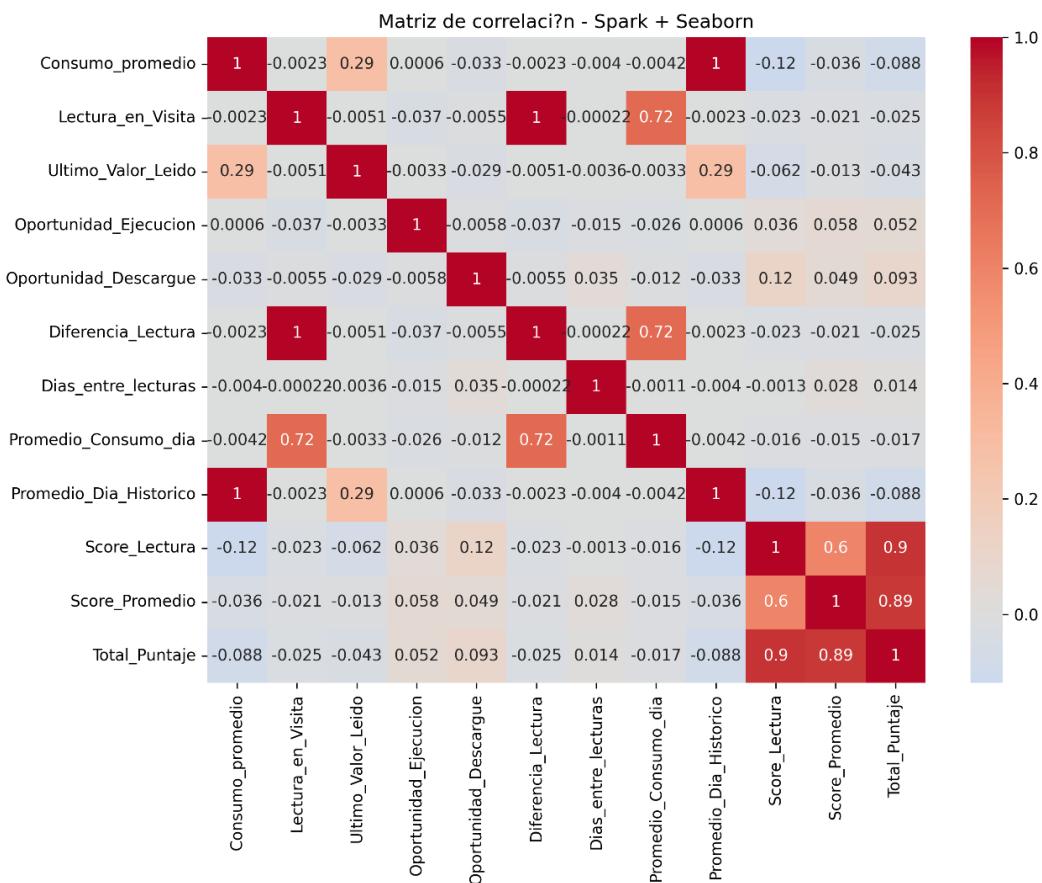
The screenshot shows the AWS Glue Crawler interface. On the left, there's a navigation sidebar with sections like AWS Glue, Data Catalog, Databases, Tables, Schemas, Connections, Crawlers, Catalogs, Catalog settings, Data Integration and ETL, and Legacy pages. The 'Crawlers' section is selected. The main area is titled 'Crawlers (1)' and contains a table with one row. The table columns are Name, State, Schedule, Last run, Last run timestamp, Log, and Table changes from last run. The single row shows 'cataloger' in the 'Ready' state, with a green 'Succeeded' icon. At the top right of the main area, there are buttons for 'Actions', 'Run', and 'Create crawler'. Below the table, it says 'View and manage all available crawlers.' and 'Last updated (UTC) May 27, 2025 at 01:05:13'.

La tabla que se almacenó en la zona trusted se catalogó en AWS Glue y se muestra dentro de la base de datos proyecto1db. Mediante SparkSession habilitado con soporte Hive, se realizan consultas exploratorias con SparkSQL para ver los tipos de columnas, sus nombres exactos y reemplazar algunos nombres para mayor claridad. Se realiza un EDA vía SparkSQL y se almacena en S3. Se sacan estadísticas descriptivas básicas, tablas de frecuencia y correlaciones.

Al analizar la proporción de valores nulos en las doce métricas numéricas evaluadas, se observa que varias presentan niveles de ausencia cercanos al 30 %, lo que demanda decisiones críticas en cuanto a imputación o eliminación. Específicamente, “Consumo\_promedio” tiene un 21,3 % de datos faltantes, “Lectura\_en\_Visita” un 19,1 %, “Último\_Valor\_Leído” un 11,3 %, “Oportunidad\_Ejecución” un 18,5 %, “Oportunidad\_Descargue” un 18,6 %, “Diferencia\_Lectura” un 28,1 %, “Días\_entre\_lecturas” un 28,7 %, “Promedio\_Consumo\_día” un 29,2 %, “Promedio\_Día\_Histórico” un 21,8 %, “Score\_Lectura” un 29,1 %, “Score\_Promedio” un 29,2 % y “Total\_Puntaje” un 29,2 %. Este nivel de ausentismo sugiere que, para variables con distribuciones sesgadas, se puede imputar usando la mediana, mientras que las filas con múltiples valores faltantes podrían eliminarse si la imputación no garantiza fiabilidad.

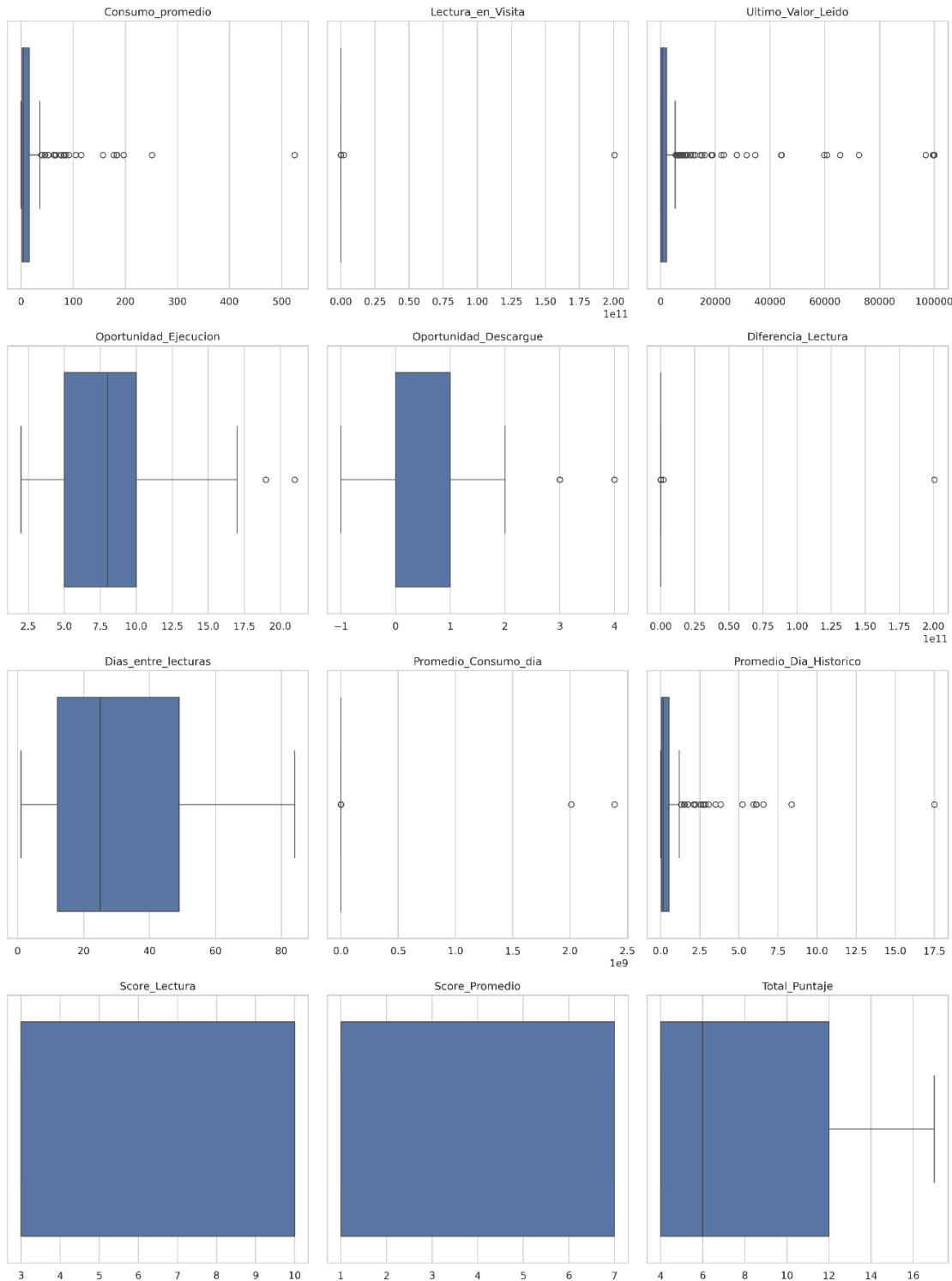
La tabla de estadísticas descriptivas revela que el número de observaciones válidas varía entre 2.514 y 3.148, lo cual refleja las proporciones de datos completos tras la tipificación explícita. “Consumo\_promedio” muestra una media de 21,3, desviación estándar de 100,7, y mediana de 5 con cuartiles en [2, 13], indicando que la mayoría de los registros reflejan consumos bajos con presencia de valores extremos. “Lectura\_en\_Visita” y “Último\_Valor\_Leído” presentan medias elevadas ( $1,54 \times 10^8$  y 6.839 respectivamente), con medianas más bajas (392 y 344), lo que

evidencia una distribución sesgada con casos extremos. En cuanto a las variables temporales, “Oportunidad\_Ejecución” tiene mediana en 8 días (Q1=5, Q3=10) y “Oportunidad\_Descargue” tiende a 0 días en la mayoría de los casos. “Diferencia\_Lectura” y “Días\_entre\_lecturas” exhiben amplitud intercuartílica ([0–34] y [10–42]) con medianas de 5 y 21 días. Finalmente, los indicadores de desempeño —“Score\_Lectura”, “Score\_Promedio” y “Total\_Puntaje”— tienen valores centrales de 3, 1 y 4 respectivamente, lo que sugiere que la mayoría de los casos se concentran en evaluaciones bajas, aunque existen algunos registros con puntajes significativamente altos (hasta 14 en lectura y 17 en puntaje total).



Al inspeccionar el heatmap, Promedio\_Dia\_Historico y Consumo\_promedio muestran correlación prácticamente unitaria, por un lado, Score\_Lectura, Score\_Promedio y Total\_Puntaje forman un bloque muy correlacionado (por encima de 0,8) al aplicar la misma lógica de cálculo. Promedio\_Consumo\_dia presenta correlaciones moderadas ( $\approx 0,7$ ) con Lectura\_en\_Visita y Dias\_entre\_lecturas, lo que tiene sentido porque más visitas o un intervalo mayor de días suelen

traducirse en consumos diarios superiores; el resto de los pares de variables exhibe correlaciones bajas (por debajo de 0,3), indicando que esas métricas aportan información más independiente.



Al recorrer los boxplots observamos que Consumo\_promedio, Promedio\_Dia\_Historico y Ultimo\_Valor\_Leido presentan distribuciones fuertemente asimétricas con medianas muy cercanas a cero y multitud de valores atípicos que alcanzan centenas o incluso miles de veces el cuartil superior, lo cual refleja picos excepcionales de consumo o lecturas extraordinarias, las variables de tiempo como Oportunidad\_Ejecucion, Oportunidad\_Descargue, Diferencia\_Lectura y Dias\_entre\_lecturas muestran rangos intercuartílicos más homogéneos alrededor de sus medianas (8, 0, 5 y 21 días respectivamente) pero también con algunos outliers que indican casos de retrasos o esperas muy prolongadas, Promedio\_Consumo\_dia mantiene la mayoría de sus observaciones por debajo de 2 unidades diarias, y en los puntajes Score\_Lectura, Score\_Promedio y Total\_Puntaje los boxplots revelan una concentración central entre 3–10, 1–7 y 4–12 con escasos valores fuera de rango, estos patrones de dispersión y presencia de outliers nos dan ya indicios claros de qué variables pueden requerir transformaciones.

Finalizando, se complementa el análisis exploratorio de las variables con eliminación de variables categóricas redundantes mediante pruebas Chi-cuadrado, se eliminan variables numéricas altamente correlacionadas según el coeficiente de Pearson, se aplican funciones personalizadas (UDF) para reclasificar la variable objetivo y se seleccionan las variables finales para consolidar un DataFrame limpio y listo para el entrenamiento del modelo.

También, aplicamos un criterio más laxo de eliminación de nulos, descartando únicamente aquellas filas que tenían todas las variables numéricas en null y, sobre ese resultado, eliminando las filas que mostraban todas las variables categóricas nulas o vacías. Los resultados fueron los siguientes, partiendo de 3 550 registros:

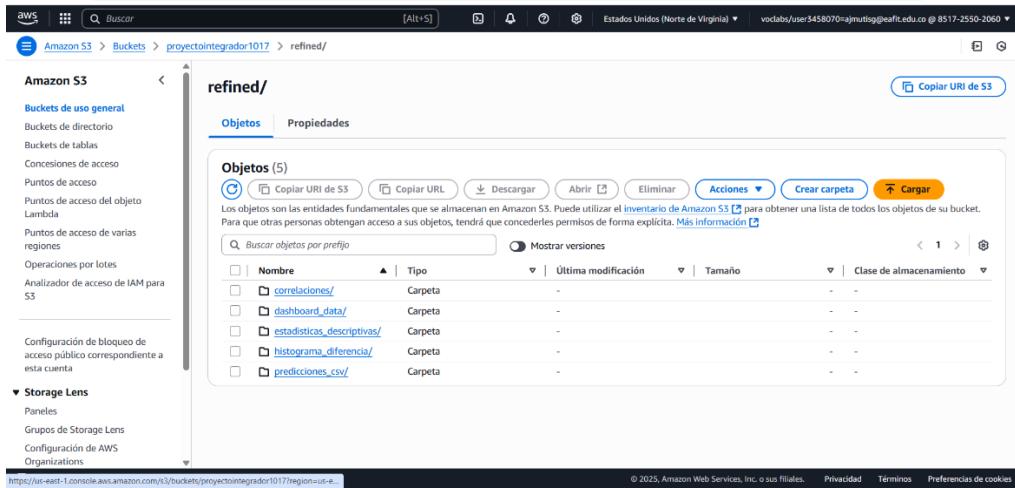
- Tras eliminar filas con todas las numéricas nulas quedaron 3 150 registros, es decir, se redujo el 11,3 % del total
- Al aplicar luego el filtro de todas las categóricas nulas o vacías no se eliminó ningún registro adicional, manteniéndose en 3 150 casos
- En conjunto conservamos el 88,7 % de los datos originales, lo que confirma que son pocos los casos con ausencia completa tanto de métricas como de categorías

Dado lo anterior, se entrenó un modelo de clasificación con LogisticRegression, integrado en un pipeline de SparkML conformado por múltiples etapas de preparación de datos y clasificación supervisada. Inicialmente, las variables categóricas fueron indexadas y codificadas en vectores binarios, mientras que las variables numéricas fueron escaladas para normalizar su rango. Paralelamente, el texto preprocessado fue convertido en vectores utilizando bigramas y CountVectorizer. Todas estas características fueron ensambladas en un solo vector unificado que sirvió como entrada para un modelo de regresión logística multiclas. La división del conjunto en entrenamiento (80%) y prueba (20%) permitió evaluar el rendimiento del modelo utilizando métricas de exactitud (accuracy) y F1 score.

Como resultado, el modelo entrenado tuvo valores de 0.8964 de exactitud (accuracy) y 0.8149 en F1 score. Los resultados son predicciones en el que según el texto observado se cataloga como anomalía o normal. Los resultados obtenidos en las métricas de evaluación (precisión y F1 score) reflejan que el modelo logra un desempeño adecuado, lo cual valida la correcta selección y preparación de las variables utilizadas. En particular, el uso conjunto de variables numéricas y categóricas permitió complementar la información objetiva, derivada de registros operativos como consumos, tiempos o puntajes, con dimensiones más cualitativas como la actividad económica, que contextualizan el comportamiento observado.

A pesar de tratarse de un modelo relativamente sencillo como la regresión logística, los resultados muestran que es capaz de predecir con precisión si la clasificación realizada por el operador (evento normal, anómalo o no revisado) es coherente con los patrones subyacentes en los datos. Esto demuestra que, incluso con modelos lineales, es posible capturar relaciones significativas cuando la calidad del dataset ha sido cuidada. Además, existe un margen claro de mejora: una depuración más rigurosa de las observaciones, la ampliación del diccionario de sinónimos o la incorporación de nuevas variables semánticas podrían enriquecer la capacidad predictiva del modelo. Este caso ilustra claramente el valor de combinar datos estructurados y no estructurados dentro de arquitecturas tipo datalake, alineándose con los principios de la analítica moderna orientada a la explotación integral de múltiples fuentes de información para la toma de decisiones basada en datos.

Por último, el resultado es un DataFrame con las predicciones al que se le limpian las columnas innecesarias y se exporta a la zona refined del datalake en S3 “s3://proyectointegrador1017/refined/predicciones\_csv”.



La visualización de resultados se desarrolló a partir de los datos procesados y almacenados en la zona refined, utilizando Plotly para generar un tablero interactivo en formato HTML. Este dashboard incluye tres gráficos clave: (1) un gráfico de barras que muestra la distribución de las categorías predichas (evento\_cat), permitiendo verificar visualmente el balance de clases y la capacidad del modelo para distinguir entre eventos normales, anómalos y no revisados; (2) un gráfico de barras que presenta los cinco eventos más frecuentes según el campo evento, lo cual ayuda a identificar patrones recurrentes en el etiquetado manual de los operadores; y (3) un histograma de la longitud de las observaciones textuales (Observacion\_Ok\_str), útil para analizar la riqueza semántica y variabilidad del contenido textual. Este panel se empaquetó como un archivo HTML y fue publicado en S3 bajo la ruta “s3://proyectointegrador1017/refined/dashboard/dashboard.html”, ofreciendo un canal práctico para explorar los resultados del modelo en tiempo real y facilitar la toma de decisiones basada en datos.

**Dashboard de Predicciones (Refined)**